Expansion of Cisco Secure AI Factory with NVIDIA to Rack-Scale Architecture

Q: What did Cisco announce?

A: Cisco is expanding the Cisco Secure AI Factory with NVIDIA to deliver industry-leading, rack-scale, liquid-cooled, AI infrastructure. Cisco will offer Supermicro liquid- and air-cooled systems and support, giving customers access to rack-scale and dense GPU systems. Cisco is the first NVIDIA technology partner to deliver an NVIDIA Cloud Partner (NCP)–compliant reference architecture built on partner-developed networking systems, spanning both Cisco® Silicon One® and NVIDIA Spectrum-X Ethernet switch silicon, unified by the Cisco Nexus® One architecture, with a choice of Cisco NX-OS (the industry's most deployed data center OS) or SONiC.

To accelerate deployment, the announcement also includes Cisco Validated Infrastructure Services (CVIS)—an end-to-end solution design validation, aligned with NVIDIA's NVIS methodology, to validate the full-stack design against the reference architecture. CVIS is anchored by a dedicated large-scale engineering AI cluster in Cisco's own AI lab, where deployment tooling, performance profiles, and software releases are continuously validated.

Q: How do enterprises, neoclouds, and sovereign clouds benefit?

A: The expansion of Cisco Secure AI Factory with NVIDIA to rack-scale infrastructure enables enterprises, neoclouds, and sovereign clouds to accelerate the deployment of emerging AI use cases, such as large-scale inference, agentic AI, and trillion-parameter model training on a full-stack AI infrastructure, with integrated security and observability from Cisco. It provides the operational framework necessary to turn raw GPU power into a deployable enterprise AI stack. By utilizing common operating models and existing technical competencies, organizations can implement a trusted AI infrastructure without the complexity of new, proprietary skill requirements.

Q: How do enterprises benefit from neoclouds built on this rack-scale architecture?

A: This architecture makes neoclouds a key path to enterprise AI adoption by ensuring that any cloud built with infrastructure delivered by Cisco is enterprise-ready on day one—using the same security controls, same networking operating model, and same skills that enterprises already have on staff.

Q: What are the key technologies in the rack-scale Cisco Secure AI Factory with NVIDIA?

A: This NVIDIA Cloud Partner (NCP) Reference Architecture (RA)–compliant rack-scale Cisco Secure AI Factory with NVIDIA architecture is based on the following:

- Industry-leading Cisco AI Networking for front-end, back-end, and storage networks
- Cisco's distributed security and Splunk® observability, fused into every layer of the stack
- Supermicro's NVIDIA NVL72 rack-scale AI systems, NVIDIA HGX–, and NVIDIA MGX–based dense GPU servers, available from Cisco's authorized channel partner ecosystem
- NVIDIA AI Enterprise software
- High-performance storage and Kubernetes platforms from our ecosystem partners, also available from Cisco's authorized channel partner ecosystem

Q: When will this rack-scale architecture be available for purchase?

A: In October 2026, organizations can order the rack-scale architecture from Cisco's authorized channel partner ecosystem. This includes Cisco, NVIDIA, Supermicro, and other ecosystem partner technologies that are part of Cisco Secure AI Factory with NVIDIA.

Q: Are Supermicro dense GPU servers also offered as part of this partnership?

A: In October 2026, Supermicro dense GPU servers (both liquid-cooled and air-cooled options) will be orderable from Cisco's authorized channel partner ecosystem. These include various platforms based on NVIDIA HGX and MGX architectures.

Q: Are Cisco Unified Compute Systems™ (Cisco UCS®) still strategic for Cisco?

A: Absolutely. Cisco UCS is the foundation of Cisco's enterprise compute portfolio and a core part of Cisco AI PODs. UCS supports the broad range of workloads enterprises run every day, from virtualized applications, databases, and software-defined storage to AI data platforms and inference. Cisco Unified Edge extends that foundation into distributed environments, bringing compute and AI closer to where data is created and used. With Supermicro, Cisco extends that portfolio into rack-scale, dense GPU infrastructure for the most compute-intensive AI workloads. Together, they give customers a comprehensive compute portfolio spanning enterprise, edge, and AI workloads, at any scale.

Q: Is Cisco Silicon One–based AI networking part of the Cisco Secure AI Factory with NVIDIA?

A: Cisco is the first NVIDIA technology partner to deliver NCP-RA compliance on its own silicon: Silicon One–based N9300 systems power the front-end fabric, alongside NVIDIA Spectrum-X Ethernet–silicon based Cisco N9100 Series Switches for the back-end network fabric, unified by the Nexus One architecture with a choice of NX-OS (the industry's most deployed data center OS) or SONiC as the operating system.

Q: How does liquid-cooled networking work in this architecture?

A: Modern rack-scale systems, such as the NVIDIA NVL72, can exceed 200 kW per rack, where liquid cooling becomes a system-level requirement rather than an option. Building on Cisco's 100% liquid-cooled AI networking systems, liquid-cooled Cisco N9000 Series Switches interoperate directly with Supermicro's rack-scale, liquid-cooled compute to deliver a rack-to-fabric liquid-cooled AI factory. This removes the thermal and power constraints that previously kept trillion-parameter training, high-throughput inference, and agentic AI out of reach for all but the hyperscalers. Also, it engineers the fabric for performance efficiency alongside the compute it serves.

Q: How is the full stack managed and monitored?

A: Unified management is delivered through Cisco Cloud Control. Within Cisco Cloud Control, network management is enabled by Cisco Nexus One, and management of compute systems will be available through Cisco Intersight® and Nexus One in CY26Q4.

Q: How does this architecture address infrastructure supply chain constraints?

A: Cisco's global supply chain scale for AI networking systems, combined with Supermicro's manufacturing and global supply chain scale for rack-scale AI systems and dense GPU servers, enables timely delivery of the infrastructure globally. This strength de-risks availability at a time when demand for GPUs, memory, CPUs, and SSDs is putting pressure on how quickly AI rack-scale systems can be built and shipped.

Q: How is the support going to be handled?

A: Cisco manages L0 and L1 ticket triage for incoming requests and routes to the appropriate solution partner. Each vendor provides support for their respective components, ensuring coordinated resolution across the full stack.

Q: What is Cisco Validated Infrastructure Services (CVIS) and how will it accelerate NCP-RA compliant deployment?

A: Cisco Validated Infrastructure Services (CVIS) is an end-to-end architecture qualification and validation program for reference architecture compliance starting with NCP-RA. CVIS includes an automation toolkit that reduces deployment and validation timelines from months to weeks. It takes the AI cluster from discovery, design, and deployment to delivery in order to accelerate time to first token.

How it accelerates deployment:

- **Qualified full-stack designs** – Every deployment starts from a proven, NCP-RA aligned reference architecture, eliminating redesign risk.
- **Repeatable provisioning and automated validation** – Every deployed configuration is verified against the reference architecture metrics such as infrastructure performance, job completion time, and token-throughput targets, using CVIS automation toolkit.
- **Specialist-assisted implementation** – This is delivered through the CVIS team with Cisco's authorized channel partners.

What customers receive:

- Fully validated cluster, compliant with NCP-RA
- Evidence report: a comprehensive performance and compliance report including validated configuration and test results

Q: How does CVIS relate to NVIDIA Infrastructure Services (NVIS)?

A: CVIS is built-on NVIS's methodology and brings the same deployment discipline from Cisco. NVIS is delivered directly by NVIDIA's infrastructure specialists, while CVIS specialists scale that model through Cisco and its authorized partners. As a result, AI cluster deployments are compliant with reference architectures supported by Cisco including Enterprise Reference Architecture (ERA), Cisco Cloud Reference Architecture (CRA), and NCP-RA.

Q: What supply chain, Return Material Authorization (RMA), and support options are available?

A: For Cisco products, its world-renowned supply chain offers diversity across build, stock, ship, and support. Customers can choose from a range of RMA options, including next-business day and 24x7x4, along with return-to-factory (RTF) service and warranty coverage. Supermicro's manufacturing scale, supply chain, and RMA options for their products give organizations the confidence that the compute they need will be delivered and supported on their timeline. Cisco manages L0 and L1 ticket triage for incoming requests, and routes to Supermicro as needed. Supermicro provides support for their products, ensuring coordinated resolution across the full stack.

C67-6678802-00 08/26